

METHOD AND APPARATUS FOR THREE-DIMENSIONAL OBJECT SEGMENTATION.

Field of the Invention

5 The present invention relates to automated vision systems, and more particularly to a system for three-dimensional object segmentation.

Background of the Invention

10 Passive techniques of stereopsis involve triangulation of features viewed from different positions or at different times, under ambient lighting conditions, as described in "Structure From Stereo--A Review," Dhond, Umesh R, and Aggarwal, J.K., IEEE Transactions On Systems, Man, And Cybernetics, Vol. 19, No, 6, Nov/Dec 1989. The major steps in stereopsis are preprocessing, matching, and recovering depth information. As described in the reference, the
15 process of matching features between multiple images is perhaps the most critical stage of stereopsis. This step is also called the correspondence problem.

It is also well known that stereo matching using edge segments, rather than individual points, provides increased immunity from the effects of isolated points, and provides an additional disambiguating constraint in matching segments of different stereoscopic images
20 taken of the same scene. A variety of algorithms can be used for matching edge segments that meet criteria for 3-D segments occurring along a smooth surface. In addition, a trinocular camera arrangement provides further information that can improve a binocular depth map with points (or edges) matched if they satisfy additional geometric constraints, such as length and orientation.

25 Once the segmented points have been identified and the depth information recovered, the 3-D object structure can be obtained which can then be used in 3-D object recognition. The purpose of this embodiment is more to segment the 3-D scene into 3-D objects that are spatially separated in a 2-D plane, rather than object recognition. Therefore, an elaborate 3-D object reconstruction is not necessary.

30 However, the prior combinations of feature detection, matching, 3-D segmentation are computationally intensive, either decreasing speed or increasing cost of automated systems.

Furthermore, prior methods lack robustness because of susceptibility to noise and confusion among match candidates. 3-D data is mostly used for object recognition, as opposed to segmentation of objects placed in a plane in 3-D space. Known techniques, typically using 2D segmentation, assume a fixed relationship between the camera system and the plane under consideration, that is, they do not facilitate specifying any arbitrary plane.

Summary of the Invention

The present invention provides a three-dimensional (3-D) machine-vision object-segmentation solution involving a method and apparatus for performing high-integrity, high efficiency machine vision. The machine vision segmentation solution converts stereo sets of two-dimensional video pixel data into 3-D point data that is then segmented into discrete objects, and subsequent characterization of a specific 3-D object, objects, or an area within view of a stereoscopic camera. Once the segmented points have been identified and the depth information recovered the 3-D object structure can be obtained which can then be used in 3-D object recognition.

According to the invention, the 3-D machine-vision segmentation solution includes an image acquisition device such as two or more video cameras, or digital cameras, arranged to view a target scene stereoscopically. The cameras pass the resulting multiple video output signals to a computer for further processing. The multiple video output signals are connected to the input of a video processor adapted to accept the video signals, such as a "frame grabber" subsystem. Video images from each camera are then synchronously sampled, captured, and stored in a memory associated with a data processor (e.g., a general purpose processor). The digitized image in the form of pixel information can then be accessed, archived, manipulated and otherwise processed in accordance with capabilities of the vision system. The digitized images are accessed from the memory and processed according to the invention, under control of a computer program. The results of the processing are then stored in the memory, or may be used to activate other processes and apparatus adapted for the purpose of taking further action, depending upon the application of the invention.

In further accord with the invention, the 3-D machine-vision segmentation solution method and apparatus includes a process and structure for converting a plurality of two-

dimensional images into clusters of three-dimensional points and edges associated with boundaries of objects in the target scene. A set of two-dimensional images is captured, filtered, and processed for edge detection. The filtering and edge detection are performed separately for the image corresponding to each separate camera, resulting in a plurality of sets of features and chains of edges (edgelets), characterized by location, size, and angle. The plurality is then subdivided into stereoscopic pairs for further processing, i.e., Right/Left, and Top/Right.

The stereoscopic sets of features and chains are then pair-wise processed according to the stereo correspondence problem, matching features from the right image to the left image, resulting in a set of horizontal disparities, and matching features from the right image to the top image, resulting in a set of vertical disparities. The robust matching process involves measuring the strength and orientation of edgelets, tempered by a smoothness constraint, and followed by an iterative uniqueness process.

Further according to the invention, the multiple (i.e., horizontal and vertical) sets of results are then merged (i.e., multiplexed) into a single consolidated output, according to the orientation of each identified feature and a pre-selected threshold value. Processing of the consolidated output then proceeds using factors such as the known camera geometry to determine a single set of 3-D points. The set of 3-D points is then further processed into a set of 3-D objects through a "clustering" algorithm which segments the data into distinct 3-D objects. The output can be quantified as either a 3-D location of the boundary points of each object within view, or segmented into distinct 3-D objects in the scene where each object contains a mutually exclusive subset of the 3-D boundary points output by the stereo algorithm.

Machine vision systems effecting processing according to the invention can provide, among other things, an automated capability for performing diverse inspection, location, measurement, alignment and scanning tasks. The present invention provides segmentation of objects placed in a plane in 3-D space. The criterion for segmentation into distinct objects is that the minimum distance between the objects along that plane (2D distance) exceed a preset spacing threshold. The potential applications involve segmenting images of vehicles in a road, machinery placed in a factory floor, or objects placed on a table. Features of the present invention include the ability to generate a wide variety of real-time 3-D information about 3-D objects in the viewed area. Using the system according to the invention, distance from one

object to another can be calculated, and the distance of the objects from the camera can also be computed.

According to the present invention a high accuracy feature detector is implemented, using chain-based correspondence matching. The invention adopts a 3-camera approach and a novel method for merging disparities based on angle differences detected by the multiple cameras. Furthermore, a fast chain-based clustering method is used for segmentation of 3-D objects from 3-D point data on any arbitrary plane. The clustering method is also more robust (less susceptible to false images) because object shadows are ignored.

Brief Description of the Drawings

These and other features of the present invention will be better understood in view of the following detailed description taken in conjunction with the drawings, in which:

Fig.1 is a functional block diagram of a 3-D object segmentation system, according to the invention;

Fig.2 is an illustration of a trinocular camera arrangement adapted for use in acquiring images for processing according to the invention; and

Fig. 3 is a flow diagram illustrating the processing of video images according to the invention.

Detailed Description

A vision system implemented in an illustrative embodiment according to the invention is illustrated in Fig. 1. The system acquires an image set from at least three cameras, performs edge processing for each independent image, performs stereoscopic correspondence and matching for pairs of images, merges the sets of stereoscopic data, performs 3-D computations based upon known camera geometry to determine 3-D features, and then clusters 3-D points into distinct objects.

The illustrative embodiment incorporates an image acquisition device 101, comprising at least three cameras 10a, 10b, 10c such as the Triclops model available from Point Grey

Research, Vancouver B.C. The cameras 10 send a video signal via signal cables 12 to a video processor 14. The three cameras are each focused on a scene 32 to be processed for objects. The video processor 14 includes a video image frame capture device 18, image processor 26, and results processor 30, all of which are connected to a memory device 22. Generally, digitized video image sets 20 from the video image capture device 18, such as a 8100 Multichannel Frame Grabber available from Cognex Corp, Natick, MA, or other similar device, are stored into the memory device 22. The image processor 26, implemented in this illustrative embodiment on a general-purpose computer, receives the stored, digitized, video image sets 24 and generates 3-D object data 28. The 3-D data 28 is delivered to the results processor 30 which generates results data dependent upon the application, and may indicate for example that the object has come too close to the camera-carrying device.

The image acquisition device 101 in the illustrative embodiment comprises an arrangement, as illustrated in Fig. 2, for acquiring image information. In the illustrative arrangement, three cameras: a right camera 222, a left camera 224, and a top camera 226 are mounted on an L-shaped support 220, with two of the cameras, the left camera 222 and the right camera 224 side-by-side, forming a line, and the third, top camera 226 mounted out of line with the other two 222, 224.

Fig. 3 provides an overview of operation according to the invention. Referring now to Fig. 3, in a first step 300, a plurality of video image signals are captured in a way that the image from each camera 222, 224, 226 is captured at substantially the same instant. This synchronization can be accomplished by having the video image frame capture device 18 send a timing or synchronization signal to each camera 222, 224, 226, or one camera may act as a master and generate a timing or synchronization signal to the others. The video signals from the image acquisition device 101 are digitized by the video image frame capture device 18, and stored into the memory device 22 for further processing. The video image frame capture device 18 includes digitizing circuitry to capture the video image input from the image acquisition device 101 and convert it at a high resolution to produce a digital image representing the two-dimensional scanned video image as a digital data set. Each data element in the data set represents the light intensity for each corresponding picture element (pixel). The digital data set generated from each camera 222, 224, 226 is stored in memory 22.

The next step 302 is to process the independent images to detect edges. In further accord with the invention, the filtering and edge detection are performed separately for the image corresponding to each separate camera, resulting in a plurality of sets of objects (or features, used interchangeably) characterized by location, size, and angle. Furthermore, features are
5 organized in the form of chains of connected edgelets. This process is based upon parabolic smoothing followed by a non-integral sub-sampling (at a specific granularity), Sobel Edge Detection, followed by True peak detection and finally chaining. This results in a list of connected edgelets (chains). Edges are defined by their position (xy) co-ordinate, magnitude and direction (orientation angle). Only features that belong to chains longer than a predetermined
10 length are passed to the next stage.

The stereoscopic sets of features and chains are then pair-wise processed according to the stereo correspondence problem, matching features from the right image to the left image 304RL, resulting in a set of horizontal disparities, and matching features from the right image to the top image, 304RT resulting in a set of vertical disparities.

15 The algorithm used here is a modified version of the algorithm presented in "A Stereo correspondence algorithm using a disparity gradient constraint" by S.B.Pollard, J.E.W. Mayhew and J.P. Frisby in Perception, 14:449-470, 1985. The modifications done are to exploit the fact that the features are connected into chains, therefore compatibility of correspondences is enforced between chain neighbors and not an arbitrary neighborhood. This is not only faster but
20 is more meaningful and robust as the neighboring points in the chains more often than not correspond to neighboring points on the 3-D object, where the disparity gradient constraint is enforced.

With regard to the disparity gradient itself, each correspondence or match-pair consists of a point in image 1 and a point in image 2 corresponding to the same point in the object. The
25 disparity vector is the vector between the points in the two images. The disparity gradient is defined between two points on the object or correspondences (or match-pairs) and it is the ratio of the difference between disparities to the average distance between the points in image 1 and image 2.

This disparity gradient constraint, which is an extension of the smoothness constraints
30 and surface-continuity constraints, sets an upper limit on the allowable disparity gradients. In theory, the disparity gradient that exists between correct matches will be very small everywhere.

Imposing such a limit provides a suitable balance between the twin requirements of having the power necessary to disambiguate and the ability to deal with a wide range of surfaces.

The algorithm itself works as follows. The initial set of possible matches for each feature is constrained using the epipolar constraint. The epipolar constraint means that for a given point in an image, the possible matches in image 2 lie on a line. The epipolar assumption is symmetric in the sense that for a point on image 2, the possible matches lie on a line in image 1. Therefore, the dimension of the search space has been reduced from two dimensions to one dimension. A potential match between a feature in the first image and a feature in the second image is then characterized by a initial strength of match (SOM). The SOM is calculated by comparing the magnitude and the direction of the edgelets that make up the features. The only matches considered are those which have a minimum amount of initial strength. Next, the disparity constraint is imposed. This step involves updating the SOM of each potential correspondence (match pair) by comparing it with the potential correspondences of the neighbors in the chains to which the features belong.

Next, a winner-take-all procedure is used to enforce uniqueness, which means that each point in image 1 can correspond to one, and only one, point in image 2 and vice-versa. The SOM for each match is compared to the SOMs of the other possible matches with the two features that are involved and only the strongest SOM is accepted. Then because of the uniqueness constraint, all other associated matches with the two features are eliminated from further consideration. This allows further matches to be selected as correct, provided they have the highest strength for both constituent features. So the above winner-take-all procedure is repeated for a fixed number of iterations.

Once the matches are obtained, the disparity vector can be obtained which is nothing but the vector between the two features. For a match between the right and left images, the disparity vector is predominantly horizontal, whereas for match between right and top images the disparity vector is predominantly vertical.

Further according to the invention, the multiple (i.e., horizontal and vertical) sets of results are then merged (i.e., multiplexed) into a single consolidated output, according to the orientation of each identified feature and a pre-selected threshold value. In an illustrative embodiment, if the orientation of a feature is between 45 and 135 degrees or between 225 and

315 degrees, then the horizontal disparities are selected; otherwise the vertical disparities are selected. The non-selected disparities data are discarded.

Processing of the consolidated output then proceeds using factors such as the known camera geometry 310 to determine a single set of 3-D features. The merged set of 3-D features is then further processed into a set of 3-D objects through a "clustering" algorithm which determines boundaries of 3-D objects.

Once the 3-D points of the features in the image are extracted they can be segmented into distinct sets, where each set corresponds to a distinct object in the scene. In this invention, the objects are constrained to lie in a known 2-D plane such as a table, ground, floor or road surface, which is typically the case. Therefore, segmenting the objects means distinguishing objects that are separated in this plane (2D distance along the plane). This procedure uses application domain information such as the segmentation plane mentioned above and a 3-D coordinate system attached to the plane. Assuming that the surface normal of this plane is the y axis (along which height is measured), this allows the selection of an arbitrary origin, x axis (along which to measure width), and z axis (along which depth is measured).

Other information that is needed for segmentation, all of which is relative to the plane coordinate system includes:

- (i) approximate range distances of the objects (z);
 - (ii) approximate lateral distance of the objects (x);
 - (iii) spacing threshold between the objects along the plane (2D distance along the xz);
- and
- (iv) approximate size, width, height, depth of the object (coordinate independent).

The first step that is performed is to convert all 3-D points to a coordinate system that is attached to the plane. Next, points are eliminated if they are too far or too close (range) or are too much to the left or right (lateral distance) and are too high (height of the object) and are too close to the plane on which they lie (xz plane). Eliminating points close to the ground plane helps remove shadows and plane-surface features. The set of all eliminated points contains points that are not given any object label.

The remaining points that do not get filtered out are then segmented into distinct object sets. Clustering is achieved by using the chain organization of the edgelets. The chains of features are broken into contiguous segments based on abrupt changes in z between successive points. This is based upon the theory that if they are contiguous in image coordinates and have similar z values then they correspond to the same object and hence the same cluster. Each of these segments now corresponds to a potentially separate cluster. Next, these clusters are merged, based on whether they overlap in x or in z . This is based upon the assumption that objects will be separated in xz . The criterion used for merging is the spacing threshold. It should be noted that, as an alternative, separate thresholds could be specified for x and z spacing.

There are several advantages of the present invention. The system provides high-accuracy edge detection, merging of disparity data from multiple views based on segment angle, chain-based segmentation; and high-speed, chain-based clustering.

Although the invention is described with respect to an identified method and apparatus for image acquisition, it should be appreciated that the invention may incorporate other data input devices, such as digital cameras, CCD cameras, video tape or laser scanning devices that provide high-resolution two-dimensional image data suitable for 3-D processing.

Similarly, it should be appreciated that the method and apparatus described herein can be implemented using specialized image processing hardware, or using general purpose processing hardware adapted for the purpose of processing data supplied by any number of image acquisition devices. Likewise, as an alternative to implementation on a general purpose computer, the processing described hereinbefore can be implemented using application specific integrated circuitry, programmable circuitry or the like.

Furthermore, although particular divisions of functions are provided among the various components identified, it should be appreciated that functions attributed to one device may be beneficially incorporated into a different or separate device. Similarly, the functional steps described herein may be modified with other suitable algorithms or processes that accomplish functions similar to those of the method and apparatus described.

Although the invention is shown and described with respect to an illustrative embodiment thereof, it should be appreciated that the foregoing and various other changes, omissions, and additions in the form and detail thereof could be implemented without changing the underlying invention.